Methods of Psychological Research Online 2000, Vol.5, No.1 Institute for Science Education
Internet: http://www.mpr-online.de

A note on the Martin-Löf test for unidimensionality

Tom Verguts and Paul De Boeck, University of Leuven[1]

Abstract

One test which is often used to investigate fit of the Rasch model to a dataset, is the Martin-Löf test for unidimensionality. This paper investigates whether (and when) its asymptotic chi-square distribution can be assumed to be appropriate. We also study the power of this test.

1. Introduction

Martin-Löf (cited in Glas & Verhelst, 1995 and Gustafsson, 1980) proposed a statistic to test for unidimensionality in a given dataset. More specifically, the statistic concerns the fit of the Rasch model, which is defined as

$$\Pr(\text{success}) = \frac{\exp(\theta - \beta)}{1 + \exp(\theta - \beta)},$$

for a person parameter θ and item (difficulty) parameter β. Under the Rasch model, this statistic is asymptotically χ^2 distributed. This note is on what "asymptotically" means in the case of the Martin-Löf statistic.

Denote I the number of items in a dataset. Martin-Löf's test consists of splitting this set in two parts (containing I_1 and I_2 items respectively) and calculating the maximum likelihood associated with the two parts. If the Rasch model holds, both sets tap the same dimension and the product of the maximum likelihoods of both parts should be approximately equal to the maximum likelihood calculated on both sets together.

---

[1] We wish to thank Norman Verhelst for his useful comments on the topic. Correspondence concerning this paper should be sent to Tom Verguts, Tiensestraat 102, 3000 Leuven, Belgium.

Formally, let t denote a score on the total test ($t = 0, \ldots, I$), and n_t denotes the number of persons attaining this score. Further, denote t_1 the score on the first and t_2 the score on the second subset respectively ($t_1 = 0, \ldots, I_1$; $t_2 = 0, \ldots, I_2$). The variable $\mathbf{t}$ denotes a combined score (t_1, t_2). The number of persons attaining the combined score $\mathbf{t}$ equals $n_{\mathbf{t}}$.

Then, the statistic is defined as $ML = -2\ln(LR)$, where LR is the likelihood ratio

$$\frac{\prod_{t} \pi_t^{n_t} \prod_{\mathrm{x}} L(\mathrm{x}|t)}{\prod_{\mathrm{t}} \pi_{\mathrm{t}}^{n_{\mathrm{t}}} \prod_{(\mathrm{x}_1,\mathrm{x}_2)} L(\mathrm{x}_1,\mathrm{x}_2|t_1,t_2)}. \qquad (1)$$

The $L(\ \ )$'s in (1) denote (conditional) likelihood functions of response patterns $\mathbf{x}$ evaluated in the conditional likelihood estimators for the item parameters of the model. The parameters π are theoretical proportions of the different scores; they are replaced by their "saturated" estimators (e.g., $\hat{\pi}_t = n_t / n$, where n denotes the total number of persons). The variables $\mathbf{x}_1$ and $\mathbf{x}_2$ denote partial response patterns, corresponding to the first and second subtest respectively. Under the null hypothesis, this statistic is χ^2 distributed with $I_1 I_2$-1 parameters (Verhelst, 1993).

The test is based on two frequency tables, one for the numerator in (1), where the number of cells equals the number of score groups, and one for the denominator in (1), where the number of cells equals $(I_1 + 1)(I_2 + 1)$, the combination of all scores in the two subtests (containing I_1 and I_2 items respectively). Since the χ^2 distribution only holds asymptotically, some rules of thumb have been proposed to assess whether a multinomial table is "full" enough to apply the χ^2 distribution (e.g., Siegel & Castellan, 1989; von Davier, 1997). We will follow von Davier, who notes that all expected frequencies should be at least equal to five (von Davier, 1997, p. 30). With this rule, if $I = 20$, and $I_1 = I_2 = I/2$, one would need at least 605 ($= 5\ (I/2 + 1)^2$) persons to satisfy this rule. This is however a minimum, since 605 persons is enough only in the trivial case that all cells have equal probability (see von Davier, 1997, p. 31). Hence, if the $t_1 \times t_2$ score table is large, even with moderately large sample sizes, the Martin-Löf test, under the null hypothesis, may not follow the stipulated χ^2 distribution.

2. Simulation study

We will illustrate the phenomenon discussed in the previous paragraph with a simulation study; the computer program used for the calculations is attached to the paper. Person abilities θ of the Rasch model are sampled from a standardnormal distribution. The number of persons (i.e., the sample size) can take on the values 500, 1000, or 5000.

The parameter vector $\boldsymbol{\beta}$ = (-1, -0.5, 0.5, 1) is taken as the building block for constructing item parameter vectors. If $I = 8$, we will take the item parameter vector to be ($\boldsymbol{\beta}$, $\boldsymbol{\beta}$); If $I = 16$, we take the item parameter vector to be ($\boldsymbol{\beta}$, $\boldsymbol{\beta}$, $\boldsymbol{\beta}$, $\boldsymbol{\beta}$), and so on, always in this order. So if $I = 16$, for example, item number 9 has a difficulty parameter β = -1. Possible values of I will be $I = 8$, 16, 24 in our study.

Suppose the model is tested by splitting the item set in two equal parts, items 1, ..., $I/2$ and $I/2 + 1$, ..., I (called a "split-half" procedure) and thus performing a Martin-Löf test. For every factor combination, 500 datasets were generated.

Results of this procedure are shown in the left part of Table 1, for different numbers of persons ("Sample size") and items (I). In this Table, we report the rejection rate in the upper half (at level α = .05), and the mean Martin-Löf values in the lower half (over all 500 replications). The theoretically expected rejection rate equals α; the expected mean Martin-Löf value is shown in the bottom row of the Table.

The Table shows that for $I = 8$, the statistic performs well for all sample sizes. However, if $I = 24$ its performance is bad and the observed statistic much too conservative (as can be observed from the zero rejection rate).

What if the number of cells in the $t_1 \times t_2$ table is lowered, for example by not splitting the item set in half but at another point? (Splitting in half results in the maximal number of cells in the $t_1 \times t_2$ table). The outcome should be a less conservative ML statistic. We investigate this by creating similar datasets as before, but splitting the item set as 1, ..., $I/4$, and $I/4 + 1$, ..., I, which is called a "split-left" procedure. The right part of Table 1 shows the result. The test performs similarly for $I = 8$ as in the split-half method. But for larger values of I, the split-left procedure does much better, although its behavior is still not very good for $I = 24$ and a relatively small sample size (of 500 or 1000 persons).

Table 1: Unidimensional data

	Procedure					
Rejection rate	Split-half			Split-left		
Sample size	$I = 8$	$I = 16$	$I = 24$	$I = 8$	$I = 16$	$I = 24$
500	.070	.010	.000	.044	.044	.004
1000	.052	.018	.000	.040	.038	.002
5000	.054	.030	.000	.068	.038	.012
Mean *ML* value						
Sample size						
500	15.627	57.343	111.359	11.314	44.887	89.572
1000	15.528	58.137	115.898	11.155	45.355	92.460
5000	15.450	60.362	124.534	11.514	46.958	96.759
Expectation	15	63	143	11	47	107

3. Power of the Martin – Löf test

Next, we investigate the power of the test by violating the model and checking the resulting rejection rate. Now, two different abilities θ_1 and θ_2 determine the responses; both have a standardnormal distribution and the correlation between θ_1 and θ_2 equals 0.40; note that this correlation is about the size one may expect in a typical dataset (e.g., Carroll, 1993, p. 92). In a first power study, the first $I/2$ items are governed by θ_1 and the last $I/2$ items by θ_2. Results of this procedure are shown in Table 2. It can be noted that if the correct splitting point is chosen (i.e., in this case, the split – half procedure), then the power is very high; the model is almost always rejected. On the other hand, if the correct splitting point is not known (the split – left procedure), the power is much lower.

Conversely, we can let the first $I/4$ items be governed by θ_1, the other items by θ_2, and look at the rejection rate. The results of this procedure are shown in Table 3. Again, if the correct splitting point is known (in this case, the split – left procedure), the power is high. On the other hand, if the splitup is incorrect, the power can be low (see left part of Table 3).

Table 2: Twodimensional data (1)

			Procedure			
Rejection rate	Split-half			Split-left		
Sample size	$I = 8$	$I = 16$	$I = 24$	$I = 8$	$I = 16$	$I = 24$
500	.974	1	1	.176	.328	.400
1000	1	1	1	.356	.768	.928
5000	1	1	1	.982	1	1
Mean *ML* value						
Sample size						
500	47.532	183.994	374.676	14.921	58.905	128.383
1000	81.173	301.037	609.441	18.182	74.896	163.873
5000	337.053	1229.360	2443.688	42.475	186.176	424.498
Expectation	15	63	143	11	47	107

Table 3: Twodimensional data (2)

			Procedure			
Rejection rate	Split-half			Split-left		
Sample size	$I = 8$	$I = 16$	$I = 24$	$I = 8$	$I = 16$	$I = 24$
500	.324	.346	.142	.958	1	1
1000	.634	.774	.822	.998	1	1
5000	1	1	1	1	1	1
Mean *ML* value						
Sample size						
500	22.528	77.141	152.178	37.529	134.619	276.328
1000	29.332	95.111	192.294	62.986	222.674	445.385
5000	83.977	230.869	459.590	267.981	913.167	1767.194
Expectation	15	63	143	11	47	107

4. Conclusion

The ML statistic is a useful one, but it will tend to be conservative if many cells in the $t_1 \times t_2$ table are low (e.g., lower than five), which is rather common; Indeed, in order to have a reliable test, the test needs to have a reasonable number of items. For 24 items, a sample size of even 5000 does not suffice. The condition of not too many low frequency cells can easily be checked. If this condition is not met, a useful alternative to the asymptotic χ^2 reference distribution might be the bootstrap procedure (von Davier, 1997). Second, the power of the test is good, but only if the correct splitup between items governed by different dimensions is made. Otherwise, very large datasets are needed to reliably detect the model violation. Hence, the investigator must have good knowledge of the content of the items involved in order to usefully apply the Martin-Löf test.

References

[1] Carroll, J. B. (1993). Human Cognitive Abilities: A survey of factor-analytic studies. New York: Cambridge University Press.

[2] Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 69-95). New York: Springer-Verlag.

[3] Gustafsson, J. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205-233.

[4] Siegel, S., & Castellan, J. N., Jr. (1989). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw – Hill.

[5] Verhelst, N. D. (1993). Itemresponstheorie. [Item response theory.] In T. J. H. M. Eggen, & P. F. Sanders (Eds.), *Psychometrie in de praktijk.* [Applied psychometrics.] Arnhem: Cito.

[6] von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data – Results of a Monte Carlo study. *Methods of Psychological Research,2,*29-48.